C02029: Doctor of Philosophy
Subject Code: 32903
February 2020

CRICOS Code: 009469A

Game theoretical adversarial deep learning algorithms for robust neural network models

Aneesh Sreevallabh Chivukula

Advanced Analytics Institute, School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

Game theoretical adversarial deep learning algorithms for robust neural network models

A thesis submitted in partial fulfilment of the requirements
for the degree of

Doctor of Philosophy
in
Analytics

by

Aneesh Sreevallabh Chivukula

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

February 2020

Abstract

Despite recent advances in deep learning, there is still a significant gap between the robustness of human perception and machine intelligence. Deep learning is not provably secure. In fact, deep neural networks are vulnerable to security attacks from malicious adversaries, which is an ongoing and critical challenge for deep learning researchers. Even innocuous perturbations in training data can change the way a deep network behaves in unintended ways. This means that imperceptibly and immeasurably small departures from the training data can result in a completely different label classification when using the model for supervised deep learning.

In this thesis, we explore adversarial deep learning algorithms. We examine how they exploit vulnerabilities in deep networks and how to make deep networks robust to their attacks. To explore the vulnerabilities, we simulate various model training processes under a range of various attack scenarios. Each attack strategy is assumed to be formulated by an intelligent adversary that is capable of either feature manipulation, label manipulation, or both. The optimal attack policy of our adversaries is determined by the solution for optimization problems that output the adversarial data. We then apply the knowledge that we learned to improve and reinforce the learning procedure so as to better defend against attacks.

As part of this research process, we developed new adversarial learning algorithms to solve for adversarial manipulations in supervised classification networks such as Convolutional Neural Networks (CNNs). The adversarial learning objective for our adversaries is to inject small changes into the data distributions, defined over positive and negative class labels, to the extent that the CNN subsequently misclassifies the data distribution. Thus, the theoretical goal of our deep learning process becomes one of determining whether a manipulation of the input data has reached a learner decision boundary, i.e., where too many positive labels have become negative labels. We began this research undertaking by first studying the performance vulnerabilities in CNNs. With these vulnerabilities identified, we were able to propose CNNs that are secure to those types of adversarial attacks.

We generate adversarial data by solving for optimal attack policies in Stackelberg games where adversaries target the misclassification performance of CNNs. In a sequential game-theoretic formulation, we model the interaction between an intelligent adversary and a deep learning model (a CNN) to generate adversarial manipulations by solving a two-player sequential noncooperative Stackelberg game where each player's payoff function increases with interactions to a local optimum. With a stochastic game-

theoretic formulation, we then extend the two-player Stackelberg game into a multiplayer Stackelberg game with stochastic payoff functions for the adversaries. Both versions of the game are resolved through the Nash equilibrium, which refers to a pair of strategies in which there is no incentive for either the learner or the adversary to deviate from their optimal strategy. In this case, the strategy pair is a learner weight and an evolutionary operation. In addition, we devised each attack scenario as a blackbox attack where the adversaries have no prior knowledge of the CNN's learning processes and its best response strategies. In each case, we show that the Nash equilibrium leads to a supervised classification network that is robust to subsequent data manipulation by a game-theoretic adversary.

We then explore adversaries who optimise variational payoff functions via data randomisation strategies on CNNs designed for multilabel classification tasks. Similarly, the outcome of these investigations is an algorithm design that solves a variable-sum two-player sequential Stackelberg game with new Nash equilibria. In designing the attack scenarios, the adversarial objective was to make small, undetectable changes to the test data. The adversary manipulates variational parameters in the input data to mislead the learning process of the CNN, so it misclassifies the original class labels as the targeted class labels. Our ideal variational adversarial manipulation is the minimum change needed to the adversarial cost function of encoded data that will result in the CNN incorrectly labelling the decoded data.

The resulting attack algorithms were a Simulated Annealing (SA) algorithm and an Alternating Least Squares (ALS) algorithm, which optimise the data manipulations by stochastically searching the strategy spaces of the variational adversaries. The optimal manipulations are found by solving the Nash equilibria optimization problem of the proposed Stackelberg game. Specifically, the optimal attack parameters in ALS and SA solve for the (variational nonlinear non-convex) adversarial cost functions that generate adversarial data. The payoff function for the variational adversary depends on the manipulations determined by a Variational Autoencoder (VAE), while the payoff function for the CNN classifier is evaluated in the input data space.

The adversarial data generated by this variant of the Stackelberg games simulates continuous interactions with the classifier's learning processes as opposed to one-time interactions. To evaluate CNN performance, we assessed the adversarial algorithms over different strategy spaces proposed for the MNIST handwritten digits database and the VGGFace2 database of human faces.

The original CNN model was then retrained using all the adversarial manipulations generated by the game-theoretic adversaries to create a secure CNN model. In an empirical demonstration, we show that the new secure CNN model is robust to subsequent game-theoretic data manipulation by adversaries. This promising result suggests that evolutionary algorithms based on game-theoretic modelling and mathematical optimization are significantly better approach to building more secure deep learning models.

We also empirically demonstrate that variational adversaries are also able to mislead CNNs. In these cases, the learning process of the CNNs was manipulated by an adversary at the input data level as well as the generated data. The optimal manipulations were

to stochastic optima in non-convex best responses strategies. We were able to encode the resulting adversarial data in terms of the multivariate statistical parameters of a Gaussian mixture model. We then retrained the original CNN on the manipulated data to give rise to a secure CNN that is robust to subsequent performance vulnerabilities from variational adversaries.

In applying this research, we developed a deep network model that discovers Granger causes in multivariate temporal financial market data. The model comprises a deep neural network (DNN) and a recurrent neural network (RNN) and discovers Granger-causal features with bivariate regression from bivariate time series data distributions. These features are subsequently used to discover Granger-causal graphs for multivariate regression on multivariate time series data distributions. Our supervised feature learning process with these proposed deep regression networks returned favourable F-tests for feature selection and t-tests against comparisons with other models. Moreover, a regression analysis on a set of experiments with real stock market data from Yahoo Finance demonstrates that our causal features are a significant improvement over the existing deep learning regression models in terms of minimising root mean squared error.

CERTIFICATE OF AUTHORSHIP

I, *Aneesh Chivukula* declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Computer Science*, *Faculty of Engineering & Information Technology* at the University of Technology Sydney, Australia, is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note: Signature removed prior to publication.

[Aneesh Sreevallabh Chivukula]

DATE: 10th February, 2020

PLACE: Sydney, Australia

Dedication

To my parents . . .

कारणाभावात् कार्याभावः ।

The absence of the effect is due to the absence of the cause.

Vaisesika Sutra 1.2.1 enunciating the Causality Principle in Indian Philosophy.

ACKNOWLEDGMENTS

I want to thank my principal supervisor, Dr. Wei Liu for his strong support and guidance throughout my degree. His scientific expertise and continuous encouragement of my research have been essential for my achievements. I also want to thank my co-supervisor, Dr. Jun Li who organized a friendly research environment and supported my casual academics at UTS.

I would like to thank Capital Markets CRC Limited (now RoZetta Institute) for providing me with a PhD scholarship. I would like to thank Australian Mathematical Sciences Institute for providing me with a postgraduate research internship.

Finally, I want to thank my wife for her support throughout my PhD.

I acknowledge Chandranath Adak (UTS) for providing this thesis template.

List of Publications

Related to the Thesis :

1. A. S. Chivukula and W. Liu, "Adversarial learning games with deep learning models," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 2758-2767. doi: 10.1109/IJCNN.2017.7966196

2. A. S. Chivukula and W. Liu, "Adversarial Deep Learning Models with Multiple Adversaries," in IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 6, pp. 1066-1079, 1 June 2019. doi: 10.1109/TKDE.2018.2851247

3. Chivukula A.S., Li J., Liu W. (2018) Discovering Granger-Causal Features from Deep Learning Networks. In: Mitrovic T., Xue B., Li X. (eds) AI 2018: Advances in Artificial Intelligence. AI 2018. Lecture Notes in Computer Science, vol 11320. Springer, Cham doi: 10.1007/978-3-030-03991-2_62

4. A. S. Chivukula, X. Yang and W. Liu, "Adversarial Deep Learning with Stackelberg Games," accepted for presentation at the 26th International Conference on Neural Information Processing (ICONIP 2019) of the Asia-Pacific Neural Network Society.

5. A. S. Chivukula, X. Yang and W. Liu, "Game Theoretical Adversarial Deep Learning with Variational Adversaries," in IEEE Transactions on Knowledge and Data Engineering (TKDE), doi: 10.1109/TKDE.2020.2972320.

Others :

6. Prabhu, C.S.R., Sreevallabh Chivukula, A., Mogadala, A., Ghosh, R., Livingston, L.M.J. 2019. Big Data Analytics: Systems, Algorithms, Applications. Springer Nature Singapore Pte Ltd.

Table of Contents

LIST OF FIGURES

FIGURE **Page**

List of Tables